Generating inflection variants of multi-word terms for French and German

Clematide, S ; Roth, L

Abstract: We describe a free Web-based service for the inflection of single words and multi-word terms for French and German. Its primary purpose is to provide glossary authors (instructors or students) of an open electronic learning management system with a practical way to add inflected variants for their glossary entries. The necessary morpho-syntactic processing for analysis and generation is implemented by finite-state transducers and a unification-based grammar framework in a declarative and principled way. The techniques required for German and French terms cover two typological different types of term creation and both can be easily transferred to other languages.

Originally published at:
Clematide, S; Roth, L (2011). Generating inflection variants of multi-word terms for French and German. In: Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011, Hamburg, Germany, 28 September 2011 - 30 September 2011. Universität Hamburg, 33-37.

Generating Inflection Variants of Multi-Word Terms for French and German

Simon Clematide, Luzia Roth
Institute of Computational Linguistics, University of Zurich
Binzmühlestr. 14, 8050 Zürich
E-mail: simon.clematide@uzh.ch, luzia.roth@access.uzh.ch

Abstract

We describe a free Web-based service for the inflection of single words and multi-word terms for French and German. Its primary purpose is to provide glossary authors (instructors or students) of an open electronic learning management system with a practical way to add inflected variants for their glossary entries. The necessary morpho-syntactic processing for analysis and generation is implemented by finite-state transducers and a unification-based grammar framework in a declarative and principled way. The techniques required for German and French terms cover two typological different types of term creation and both can be easily transferred to other languages.

Keywords: morphological generation, morphological analysis, multi-word terms, syntactic analysis, syntactic generation

1. Introduction

In the age of electronic media and rapid proliferation of technical terms and concepts, the use of glossaries and their dynamic linkage into running text seems to be important and self-evident in the area of e-learning. However, depending on the morphological properties of a language, e.g. the use of compounds or multi-word terms or the degree of surface modification that inflection imposes on words, the task of constructing inflected term variants from typically uninflected glossary entries is not a trivial task.

In this article, we describe two Web services for inflected term variant generation that illustrate the different requirements regarding morphological and syntactic processing. Whereas French shows modest inflectional variation in comparison to German, French requires more effort regarding syntactic analysis of complex nominal phrases. For German, guessing the correct inflection class of unknown compounds is more important.

A linguistically informed method for inflected term variant generation involves morphological and syntactical analysis and generation. In order to ensure this bidirectional processing, declarative linguistic frameworks such as finite-state transducers and rule-based unification grammars are beneficial. For a practical system, however, one wants to be able to analyze a wider range of expressions than what should actually be generated and presented to the user, e.g. entries in the form of back-of-the-book indexes should be understood by the system, but these forms will not appear in running text.

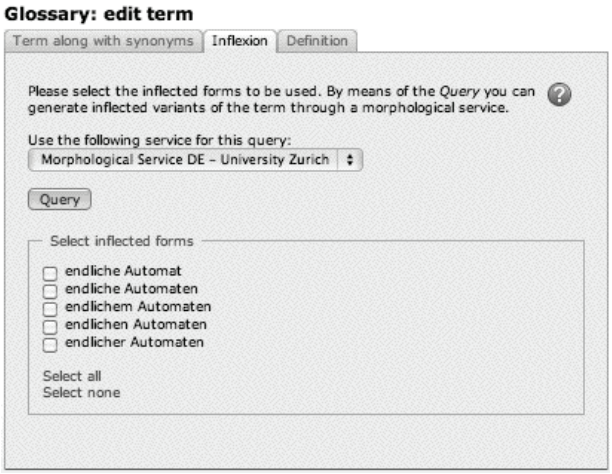

Figure 1: Screenshot of the glossary author interface

The main application domain for our services is the e-Learning Management Framework OLAT[1] where we provide glossary authors with an easy but fully controllable way to add inflected variants for their glossary entries. Our free Web-based generation service[2] is only called once for a given term, viz. when the

[1] See http://www.olat.org for further information about the open source project OLAT (Online Learning and Training).

[2] The service is realized as a Common Gateway Interface (CGI), and it delivers a simple XML document customized for further processing in the glossary back-end of the e-learning

glossary author edits an entry. As shown in Fig. 1, the glossary author is free to select or deselect any of the generated word forms.

2. Methods and Resources

In this section, we first describe the lexical and morphological resources used for French and German. In section 2.2 we discuss the implementation of the syntactic processing module.

2.1. Lexical Resources

2.1.1. Lexical resources for French

Morphalou[3], a lexicon for inflected word forms in French (95,810 lemmata, 524,725 inflected forms), was used as a lexical resource to automatically build the finite-state transducer [4] which provides all lexical information, including word forms and morphological tags.

After the first evaluation ofour development set, some modifications were made to extend the vocabulary: As derivations with neo-classical elements are quite common in terminological expressions, all adjectives were additionally combined with the prefixes of a list[5] to create derivational forms such as *audiovisuel, interethnique* or *biomédical.*

Furthermore, from all lexicon entries containing a hyphen the beginning from the entry including the hyphen was extracted. This string was taken as a prefix and combined with nouns to cover cases like *demi-charge*.

2.1.2. Lexical resources for German

We use the lexicon *molifde* (Clematide, 2008), which was mainly built by us by exploiting a full form lexicon generated by Morphy (Lezius, 2000), the German lexicon of the translation system OpenLogos [6] , and the morphological resource Morphisto (Zielinski & Simon, 2008). The manually curated resource contains roughly 40,000 lemmas (nouns, adjectives, verbs), and by applying automatic rules for derivation and conversion an additional set of 100,000 lemmas is created.

As noun compounds are the most common and productive form of terms in German, a suffix-based inflection class guesser for nouns is necessary. In an evaluation experiment with 200 randomly selected nouns from a sociology lexicon[7], about 40% of the entries were unknown. We implemented a finite-state based ending guesser by exploiting frequency counts of lemma endings (3 up to 5 characters) from our curated lexicon. Roughly 80% of the 73 unknown singular nouns got their correct inflection class. The finite-state based ending guesser is tightly coupled with the finite-state transducer derived from our lexicon. See Clematide (2009) for technical implementation details.

2.2. Morpho-syntactic Analysis and Generation

While the generation of inflected variants for single words can be easily done with the help of finite-state techniques only, this is not the case for a proper treatment of complex multi-word terms. Therefore, we decided to use a unification-based grammar framework for syntactic processing.

The Xerox Linguistic Environment (XLE) has several benefits for our purposes:

Firstly, finite-state transducers for morphological processing integrate in a seamless and efficient way. Additionally, different tokenizer transducers can be specified for analysis and generation. This proved to be useful for the treatment of French, e.g. regarding the treatment of hyphenated compounds.

Secondly, there are predefined commands in XLE for parsing a term to its functional structure, neutralizing certain morpho-syntactic features, and generating all possible strings out of an underspecified functional structure.

Thirdly, the implementation of optimality theory in XLE allows a principled way of specifying preference heuristics, for instance for the part of speech of an ambiguous word. Additionally, using optimality marks allows to analyze more constructions than what should be generated, e.g. terms in the format of back-of-the-book indexes as *Automat, endlich*. With the same technique different lexical specification conventions of French

management software OLAT. See http://kitt.cl.uzh.ch/kitt/olat.

[3] See http://www.cnrtl.fr/lexiques/morphalou for this resource, which is freely available for educational and academic purposes.

[4] We use the Xerox Finite State Tools (XFST) (Beesley & Karttunen, 2003), which seamlessly integrate with the Xerox Linguistic Environment (XLE), see http://www2.parc.com/isl/-groups/nltt/xle.

[5] http://fr.wiktionary.org/wiki/Catégorie:Préfixes_en_français

[6] Containing approx. 120,000 entries with inflection class categorizations of varying quality, see http://logos-os.dfki.de.

[7] http://www.socioweb.org

	Terms	Correct Generation	Incorrect Generation	Accuracy
Development Set	400	376	24	94%
			parse failure: 19	
			wrong parse: 5	
Test Set	50	48	2	98%
			parse failure: 1	
			wrong parse: 1	

Table 1: Evaluation results for French from the development set and test set

adjectives can be handled by the XLE grammar. Lexicon entries like *grand, e* or *grand/e* or *grand(e)* are parsed and will result in the same output *grand, grande, grands, grandes*.

Lastly, dealing with unknown words is supported in XLE in a way that parts of a multi-word term that do not undergo inflection may be analyzed and regenerated verbatim. This is useful for the treatment of postnominal prepositional phrases.

The use of a full-blown grammar engineering framework for the generation of inflected term variants might be seen as too much machinery at first sight. However, the experience we gained with this approach is definitely positive. Despite the expressivity of the framework, the processing time needed for the processing of one multi-word term is about 200ms on an AMD Opteron 2200 MHz. Given the fact that our service is only called when an entry is created by a glossary author, this performance is adequate.

2.2.1. French multi-word terms

As French is more analytic than German, compounding is less prominent. The words in a multi-word term are syntactically depending on each other and require syntactic processing. The most common construction for multi-word terms is a noun combined with a preposition and a noun phrase (e.g. *droit de l'individu*). Such constructions typically correspond to German compounds. Each noun may be modified by one or more adjectives. For a correct generation of all inflected variants, the core noun and its core adjectives have to be identified, as these are the only parts to be altered for inflected variants. The core part of a French multi-word term is typically the one preceding the preposition (e.g. *droit de l'individu* → *droits de l'individu*). Due to this fact, even terms with unknown words can be handled as long as they follow the preposition. In our XLE grammar, a default parsing strategy for unknown words occurring after a preposition is built-in and for the generation side such input is copied unchanged.

Further constructions for multi-word terms are: a noun with one or more adjectives, expressions with a hyphen (e.g. *éthylène-glycol*), noun-noun combinations (e.g. *assurance maladie*) or combinations of several nouns with *et* or *ou* (e.g. *cause et effet*). For our development set of 400 terms (see section 3.1.1 for further details), we get the following distribution: terms with prepositions (190), terms with adjectives (183), noun-noun combinations (16), terms with hyphens (9), combination of type noun *et* noun (2).

2.2.2. Preference heuristics for French

If the parsing of a one-word input term results in ambiguous structures, nouns are preferred to adjectives and verbs, as glossary entries often are nouns. For ambiguous structures of multi-word input terms the sequence noun-adjective is preferred to noun-noun, e.g. *église moderne = noun + adjective* instead of *noun + noun*. If a term is a combination of two nouns, only the first one is inflected, e.g. *assurance maladie* → *assurances maladie*.

In expressions with a hyphen, inflection is carried out by treating the hyphenated part of the term as normal word: Core adjectives or nouns with a hyphen are inflected, all others are not, e.g. *éthylène-glycol* → *éthylène-glycols*, or *document quasi-négociable* → *documents quasi-négociables*. In these two examples, the second part of the hyphenated expression is a core noun and has to be inflected. But there are cases where both parts of the hyphenated expression are non-core nouns. They are not inflected as in the example *égalité homme-femme* → *égalités homme-femme*. This example follows the

construction of a noun-noun multi-word term and is treated as such.

2.2.3. German multi-word terms

A detailed technical report on the XLE-based generation and analysis part for German can be found in Clematide (2009). Currently, German multi-word terms are restricted to the combination of an attributive adjective and a noun that may be given in the textual form of 'adjective noun' or as back-of-the-book index entry 'noun, adjective'. For instance, the lexicon entry *endlicher Automat* (finite state automaton) leads to the following 6 inflected forms: *endlichem Automaten, endlicher Automat, endlicher Automaten, endlichen Automaten, endliche Automat, endliche Automate.*

2.2.4. Related work

As far as term structures in French are concerned, Daille (2003) gives an overview that provided a base for our own analysis of multi-word terms structures. This classification was adapted and extended according to our potential glossary entries.

Jacquemin (2001) developed FASTR, a system for identifying morphological and syntactical term variants for French and English where also minor lexical modifications may take place. We did not use this system mainly for two reasons: we also had to treat German and the creation of lexical variants was of minor importance for us too.

In her contrastive study, Savary (2008) discusses different approaches of computational inflection regarding multi-word units. She emphasizes the lexical and sometimes idiosyncratic nature of multi-word expressions that may lead to problems for simple rule-based syntactic systems. However, our small-scale evaluation presented in the next section does not indicate severe problems for our approach.

3. Evaluation

In this section, we present results of our tools derived from two small-scale evaluations.

3.1.1. French

A development set with 400 and a test set with 50 glossary entries were taken randomly from *EuroVoc*[8], the EU's multilingual thesaurus. Table 1 shows the results for both data sets. Parsing failures were due to unknown vocabulary entries such as abbrevations (e.g. *CEC*, *P et T*) or compounds (e.g. *désoxyribonucléique*, *spéctrométrie*). Surprisingly, quite common French words like *jetable* and *environnemental* (appeared 5 times in the development set) were not covered by the lexicon. To alleviate the problem of missing vocabulary, additional open resources may be exploited[9]. Wrong parses were caused by ambiguities between nouns and adjectives.

3.1.2. German

50 German multi-word terms were selected randomly from the preferred terms in *EuroVoc*. Without the unknown word guesser, the generation of inflected variants fails for 10 terms, resulting in an accuracy of 80%. Applying the unknown word guesser for nouns allows a correct generation in 5 cases, thus giving an accuracy of 90%. 2 cases are due to unknown short nouns (the guesser requires a minimal length), 2 cases are due to unknown adjectives, and 1 case originated from an implementation error concerning adjectival nouns as *Beamter* (civil servant).

4. Conclusions

We have built a practical morphological generation service for French and German terms based on linguistically motivated processing. For multi-word terms, more constructions can be easily added through modifications of the syntactic term grammar.

In order to achieve a higher lexical coverage, other resources can be integrated. In our French system, there is already an interface that allows for simple addition of new regular nouns and adjectives. For German, additional syntactic constructions for multi-word terms will be added.

In order to resolve ambiguities on the level of parts of speech within multi-token terms, a part-of-speech tagging approach is feasible. However, for that purpose a specifically trained tagger is necessary

[8] http://eurovoc.europa.eu/drupal

[9] E.g. wiktionaries (http://fr.wiktionary.org/wiki/Wiktionnaire), or different lexica with inflected forms such as lefff - lexique des formes fléchies du français (http://www.labri.fr/perso/-clement/lefff), Dictionnaire DELA fléchi du français (http://infolingu.univ-mlv.fr), or Lexique3 (http://www.-lexique.org), a lexicon with lemmata and grammatical categories.

In a future step, we plan to extract nominal groups from a syntactically annotated corpus and use that material for the training of a part-of-speech tagger.

5. Acknowledgements

The University of Zurich supported this work by IIL grant funds. Luzia Roth implemented the French part under the supervision of Simon Clematide. The implementation of the lexicographic interface in OLAT was realized by Roman Haag under the supervision of Florian Gnägi.

6. References

Beesley, K.R., Karttunen, L. (2003): Finite-State Morphology: Xerox Tools and Techniques. CSLI Publications.

Clematide, S. (2008): An OLIF-based open inflectional resource and yet another morphological system for German. In A. Storrer et al. (Eds.), Text Resources And Lexical Knowledge: selected papers from the 9th Conference on Natural Language Processing, KONVENS, Mouton de Gruyter, pp. 183-194.

Clematide, S. (2009): A morpho-syntactic generation service for German glossary entries. In S. Clematide, M. Klenner, and M. Volk (Eds.), Searching Answers: Festschrift in Honour of Michael Hess on the Occasion of His 60th Birthday, Münster, Germany: Monsenstein und Vannerdat, pp. 33-43.

Daille, B. (2003): Conceptual Structuring Through Term Variations. In Proceedings of the ACL 2003 workshop on multiword expressions analysis acquisition and treatment, pp. 9-16.

Jacquemin, C. (2001): Spotting and Discovering Terms through Natural Language Processing. Massachusetts Institute of Technology.

Lezius, W. (2000): Morphy - German morphology, Part-of-Speech tagging and applications. In Proceedings of the 9th EURALEX International Congress, Stuttgart, pp. 619-623.

Savary, A. (2008): Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches. Linguistic Issues in Language Technology - LiLT, 1(2).

Zielinski, A., Simon C. (2008): Morphisto: An Open-Source Morphological Analyzer for German. In Proceedings of the FSMNLP 2008, pp. 177-184.